3) Fruit Classification.

This question use the fruit data to determine which of the following fruits

- **Banana**
- **Orange**
- **Other Fruit**

are more likely to be Long, Not Sweet, and Not Yellow

Training Data

Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	Total
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other Fruit	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

Basic Probability

- Prior probabilities

P (Banana) = 0.5 = 500/1000

P (Orange) = 0.3 = 300/1000

P (Other Fruit) = 0.2 = 200/1000

- Probability of "Evidence"

P (Long) = 0.5 = 500/1000

P (Not Sweet) = 0.35 = 350/1000

P (Not Yellow) = 0.2 = 200/1000

Step 1 – Check Banana is more likely to be Long, Not Sweet, and Not Yellow.

By comparison, the formula of using **Bayes Theorem** to solve this problem is

P(Banana|Long, Not Sweet and Not Yellow)

= P(Banana|Long ∩ Not Sweet ∩ Not Yellow)

$$= \frac{P\,(\text{Long} \cap \text{Not Sweet} \cap \text{Not Yellow} \mid \text{Banana}) * P(\text{Banana})}{P\,(\text{Long} \cap \text{Not Sweet} \cap \text{Not Yellow})}$$

= (count (Long, Not Sweet, Not Yellow, Banana)/count (Banana))

* (count (Banana)/count(total))

----

count(Long, Not Not Yellow, Banana) /count(total)

Note: It is hard to find count(Long, Not Not Yellow, Banana).

Applying Naive Bayes Formula -

P(Banana|Long, Not Sweet and Not Yellow)

P(Long|Banana) * P(Not Sweet|Banana) * P(Not Yellow|Banana) * P(banana)

= ---

P(Long) * P(Not Sweet) * P(Not Yellow)

- **P(Long|Banana)** = count (Long, Banana)/count(Banana) = 400 / 500 = 0.8
- **P(Not Sweet|Banana)**

  = count (Not Sweet, Banana) / count (Banana) = 150/500 = 0.3

- **P(Not Yellow|Banana)**

  = count (Not Yellow, Banana) / count (Banana) = 50/500 = 0.1

Substitute Values in below Formula -

P(Long|Banana) * P(Not Sweet|Banana) * P(Not Yellow|Banana) * P(banana)

= ---

P(Long) * P(Not Sweet) * P(Not Yellow)

= (0.8*0.3*0.1*0.5)/ P(evidence) = **0.012/ P(evidence)**

Step 2 – Check Orange is more likely to be Long, Not Sweet, and Not Yellow.

By comparison, the formula of using **Bayes Theorem** to solve this problem is

P(Orange|Long, Not Sweet and Not Yellow)

= P(Orange |Long ∩ Not Sweet ∩ Not Yellow)

P(Long ∩ Not Sweet ∩ Not Yellow | Orange) * P(Orange)

= --

P(Long ∩ Not Sweet ∩ Not Yellow)

= (count (Long, Not Sweet, Not Yellow, Orange)/count (Orange)) * (count (Orange)/count(total))

----

count(Long, Not Not Yellow, Orange) /count(total)

Note: It is hard to find count(Long, Not Not Yellow, Orange).

Applying Naive Bayes Formula -

P(Orange |Long, Not Sweet and Not Yellow)

P(Long| Orange) * P(Not Sweet| Orange) * P(Not Yellow| Orange) * P(Orange)

= ---

P(Long) * P(Not Sweet) * P(Not Yellow)

- **P(Long| Orange)** = count (Long, Orange)/count(Orange) = 0/ 300 = 0
- **P(Not Sweet| Orange)**

  = count (Not Sweet, Orange) / count (Orange) = 150/300 = 0.5

- **P(Not Yellow| Orange)**

  = count (Not Yellow, Orange) / count (Orange) = 0/300 = 0

Substitute Values in below Formula -

P(Long| Orange) * P(Not Sweet| Orange) * P(Not Yellow| Orange) * P(Orange)

= ---

P(Long) * P(Not Sweet) * P(Not Yellow)

= (0*0.5*0*0.3)/ P(evidence) = **0**

The value of P(Orange | Long, Not Sweet and Not Yellow) was zero in the above example, because, P(Long | Orange) and P(Not Yellow| Orange) was zero. That is, there were no ‘Long’ and ‘Not Yellow’ oranges in the training data.

It makes sense, but when you have a model with many features, the entire probability will become zero because one of the feature’s value was zero.

Step 3 – Check Other Fruit is more likely to be Long, Not Sweet, and Not Yellow.

By comparison, the formula of using **Bayes Theorem** to solve this problem is

P(Other Fruit |Long, Not Sweet and Not Yellow)

= P(Other Fruit |Long ∩ Not Sweet ∩ Not Yellow)

$$= \frac{P(\text{Long} \cap \text{Not Sweet} \cap \text{Not Yellow} \mid \text{Other fruit}) * P(\text{Other Fruit})}{P(\text{Long} \cap \text{Not Sweet} \cap \text{Not Yellow})}$$

$$= \frac{(\text{count (Long, Not Sweet, Not Yellow, Other Fruit)/count (Other Fruit)}) * (\text{count (Other Fruit)/count(total)})}{\text{count(Long, Not Sweet , Not Yellow, Other Fruit) /count(total)}}$$

Note: It is hard to find count (Long, Not Sweet, Not Yellow, Other Fruit).

Applying Naive Bayes Formula -

P(Other Fruit |Long, Not Sweet and Not Yellow)

$$= \frac{P(\text{Long}| \text{Other Fruit}) * P(\text{Not Sweet}| \text{Other Fruit}) * P(\text{Not Yellow}| \text{Other Fruit}) * P(\text{Other Fruit})}{P(\text{Long}) * P(\text{Not Sweet}) * P(\text{Not Yellow})}$$

- **P(Long| Other Fruit)** = count (Long, Other Fruit)/count (Other Fruit)

=100/200 = 0.5

- **P(Not Sweet| Other Fruit)**

  = count (Not Sweet, Other Fruit) / count (Other Fruit) = 50/200 = 0.25

- **P(Not Yellow| Other Fruit)**

  = count (Not Yellow, Other Fruit) / count (Other Fruit) = 150/200 = 0.75

Substitute Values in below Formula -

P(Long| Other Fruit) * P(Not Sweet| Other Fruit) * P(Not Yellow| Other Fruit)

* P(Other Fruit)

= --

P(Long) * P(Not Sweet) * P(Not Yellow)

= (0.5*0.25*0.75*0.2) / P(evidence) = **0.01875 / P(evidence)**

The denominator is the same for all 3 cases, so it's optional to compute.

Clearly, Other Fruit gets the highest probability, so that will be our predicted class.

Summary -

By an overwhelming margin (0.01875 >> 0.012), we classify this Not Sweet/Long/Not Yellow fruit as likely to be an Other Fruit.